UNIT 2 ASSIGNMENT

Managing Your Data in ML

## Instructions

The questions below will prepare you for future interviews as they relate to concepts discussed throughout the week. You’ve practiced these concepts in the coding activities, exercises, and coding portion of the assignment. Now, let’s formulate your programming into well-thought responses.

Except as indicated, use this document to record all your assignment work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your facilitator as part of your assignment completion. You may also have additional supporting documents that you will need to submit. Your facilitator will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used or shared outside the course, you should take care to obscure any information you feel might be of a sensitive or confidential nature.

*Begin your assignment by completing the questions below. Directions to submit your work can be found on the assignment page. Information about the grading rubric is available on any of the course assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the assignment.*

Week 2 Written Portion

Building a Modeling Dataset

Answer the questions below about building a model dataset and understanding your data through analysis and visualization.

## Questions:

1. What does it mean to have a “modeling dataset”?

| A “modleing dataset” is a cleaned up set of data. That means, no missing values, no outliers, etc. |
| --- |

1. What steps would you take with a raw dataset to end up with a modeling dataset?

| First look at the data. Then, once you clean up your data to have a useable dataset, you can start making labels and correlations within the dataset. That’s when you have a modeling dataset |
| --- |

1. What is the difference between nominal data and ordinal data? Explain with an example.

| Nominal and ordinal data are both categorical data. Nominal is often strings and doesn’t have ordering; it usually uses one-hot encoding to transform the column to numbers for analysis. Ordinal does have a ranking and has binary indicator features. Ex. Nominal: blonde/brown/black hair but Ordinal: first place/second/third OR movie rating from 5-10 stars –can be numeric |
| --- |

1. Why is data visualization an important part of the data preparation process?

| Data vizualization helps you understand the skew and outliers of your raw data. This helps you determine what needs to be adjusted to ensure the data is meaningful, predicts the right things –accurately, and to clean the data from missing values through winsorization. |
| --- |

1. What is an outlier?

| Simply put, an outlier is a data point that is far from all the others. It usually is in the top 1% and skews the mean values and makes your error bars bigger. |
| --- |

1. Name a few libraries used for data analysis and visualization and explain when you would use each library.

| I would use:  NumPy: powerful arrays for bigger data to plug into scipy or pandas functions  Pandas: data frames, tables, head and new columns  Matplot: graphing, vizualization, histograms; pyplot  Scipy: for the mean and z-score, stats, etc. |
| --- |

*To submit this assignment, please refer to the instructions in the course*. 